RETO INEGI

ANALISIS DE CIENCIA DE DATOS

EQUIPO 2

MARCOS AQUINO GARCIA	A00835576
SERGIO ALEJANDRO ESPARZA GONZÁLEZ	A01625430
CARLOS ALBERTO GÓMEZ SAN PEDRO	A01658377
PEDRO SOTO JUÁREZ	A00837560
JOSÉ FRANCISCO OBREGÓN GAXIOLA	A00227502
ARATH MENDIVIL MORA	A01660670

PROFESORES

FELIPE CASTILLO RENDÓN
LUIS DANIEL MENDOZA MORALES

INEGI

ALEJANDRA FIGUEROA MARTÍNEZ

MARCOS AQUINO GARCIA

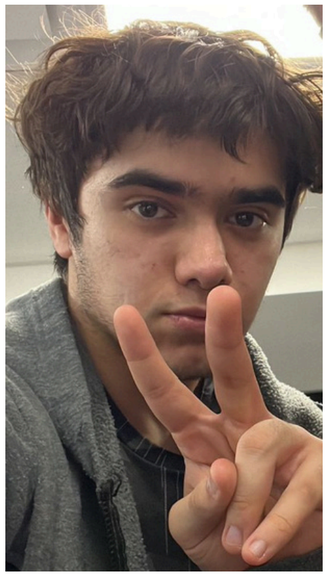
ESCRITURA CODIGO PYTHON

SERGIO ALEJANDRO ESPARZA GONZÁLEZ

ELABORACIÓN DE REPORTE

JOSÉ FRANCISCO OBREGÓN GAXIOLA

ELABORACION DE REPORTE



PEDRO SOTO JUÁREZ

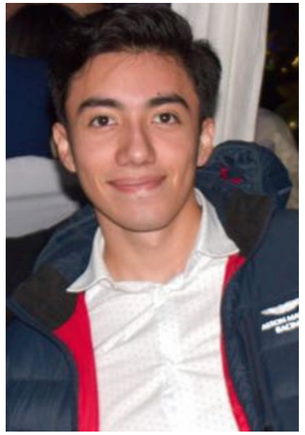
ELABORACION DE REPORTES

ARATH MENDIVIL MORA

ELABORACIÓN DEL REPORTE Y PRESENTACIÓN EJECUTIVA

CARLOS GÓMEZ SAN PEDRO

MODELO ER, ANALISIS DE MODELOS Y PRESENTACIÓN EJECUTIVA



OBJETIVO

SE ESCOGIÓ ANALIZAR EN TRES ENTIDADES DISTINTAS, QUE SERÍAN CIUDAD DE MÉXICO, SONORA Y VERACRUZ, LA CANTIDAD DE PERSONAS MAYORES DE 18 AÑOS LAS CUALES TIENEN EDUCACIÓN POST BÁSICA. ESTO NOS AYUDARÁ A PREDECIR ALGUNAS DE LAS CAUSAS POR LAS CUALES EN MÉXICO NO SE TERMINA LA EDUCACIÓN POST BÁSICA ANALIZANDO DATOS DE VIVIENDA, SALUD Y ENTRE OTROS.

CREAR UN MODELO PREDICTIVO QUE PUEDA DETERMINAR LAS CONDICIONES DE VIDA DE LA POBLACIÓN MAYOR DE 18 AÑOS EN MÉXICO QUE NO COMPLETA SU EDUCACIÓN POST BÁSICA. EL MODELO BUSCA IDENTIFICAR FACTORES QUE INFLUYAN EN ESTAS CONDICIONES, COMO EL ACCESO A SERVICIOS BÁSICOS, COMO TRANSPORTE, INTERNET, ETC. CON EL FIN DE COMPRENDER MEJOR LOS DETERMINANTES DE LA CALIDAD DE VIDA DE ESTE GRUPO Y DESARROLLAR ESTRATEGIAS EFECTIVAS PARA MEJORAR ESTAS CONDICIONES.

CRISP

CONOCIMIENTO DEL NEGOCIO

- ***Criterios de éxito:*** *Crear un modelo predictivo que alcance un óptimo de precisión, capaz de identificar con éxito los factores predictivos más significativos que contribuyen a la educación incompleta y, por tanto, afectan la calidad de vida.*
- ***Riesgos y contingencias:*** *Posibles inexactitudes en los datos del INEGI, problemas de integración de datos, y la necesidad de ajustes metodológicos debido a datos faltantes o atípicos.*

COMPRENSIÓN DE LOS DATOS DEL NEGOCIO

- ***Fuentes de datos:*** *Datos proporcionados por el INEGI Censo de Población y Vivienda, con variables como la población con educación posbásica y características demográficas básicas.*
- ***Técnicas de exploración de datos:*** *Análisis descriptivo utilizando medidas de tendencia central y dispersión, histogramas para análisis de distribución, y boxplots para identificar valores atípicos.*
- ***Calidad de datos:*** *Verificar la completitud y exactitud de los datos, manejar valores faltantes y corregir errores.*

PREPARACIÓN DE LOS DATOS

- ***Selección de datos:*** *Se hará énfasis en las variables como PSINDER (población sin acceso a servicios de salud), VPH_NDACMM (viviendas sin vehículo) y VPH_SINCINT (viviendas sin internet).*
- ***Limpieza de datos:****, Manejo de valores faltantes mediante métodos de imputación apropiados, y conversión de categorías a numérico si es necesario.*
-
- ***Integración de datos:*** *Combina datasets de diferentes entidades para un análisis consolidado.*
- ***Formateo de datos:*** *Asegurar que todos los datos estén en formatos adecuados para el análisis.*

CRISP

MODELACIÓN

- ***Técnicas de modelado:*** *Emplear algoritmos de predicción para comparar los grupos de interés. Métodos potenciales incluyen K Neighbors, árboles de decisión, o random forests.*
- ***Configuración de pruebas:*** *División de datos en conjuntos de entrenamiento y prueba para validar la efectividad del modelo.*
- ***Criterios de evaluación:*** *Utilizar la precisión del modelo, la matriz de confusión, R^2, entre otros indicadores de desempeño.*

EVALUACIÓN DE LOS MODELOS

- ***Evaluación de resultados:*** *Comparar los resultados del modelo con los objetivos y criterios de éxito del negocio. Interpretar los modelos en términos de impacto social y económico.*
- ***Revisión del proceso:*** *Revisar cada etapa del proceso CRISP-DM para asegurar que se ha ejecutado correctamente y ajustar donde sea necesario.*
- ***Pasos a seguir:*** *Decidir sobre la implementación del modelo y la formulación de recomendaciones de política basadas en los hallazgos del modelo.*

IMPLEMENTACIÓN DE LA SOLUCIÓN

1. ***Implementación***:
   - *Implementar el modelo en el entorno de producción en sistemas existentes o desarrollando una interfaz de usuario si es necesario.*
   - *Asegurarse de que el modelo esté correctamente configurado y pueda generar predicciones en tiempo real según sea necesario*
   - *Establecer un proceso de monitoreo continuo para supervisar el rendimiento del modelo en producción.*
   - *Realizar ajustes periódicos al modelo según sea necesario, como la actualización con nuevos datos o la recalibración de parámetros.*

DESCRIPCIÓN DE DATOS

1. DATOS DEL DENUE (DIRECTORIO ESTADÍSTICO NACIONAL DE UNIDADES ECONÓMICAS):

Descripción: *Incluye información sobre la identificación, tamaño, ubicación y tipo de negocio de las unidades económicas activas en México.*

2. DATOS DEL CENSO DE POBLACIÓN Y VIVIENDA:

Descripción: *Proporciona datos sobre la dimensión, estructura y distribución de la población, así como características sociodemográficas y culturales de los habitantes y detalles de las viviendas en México.*

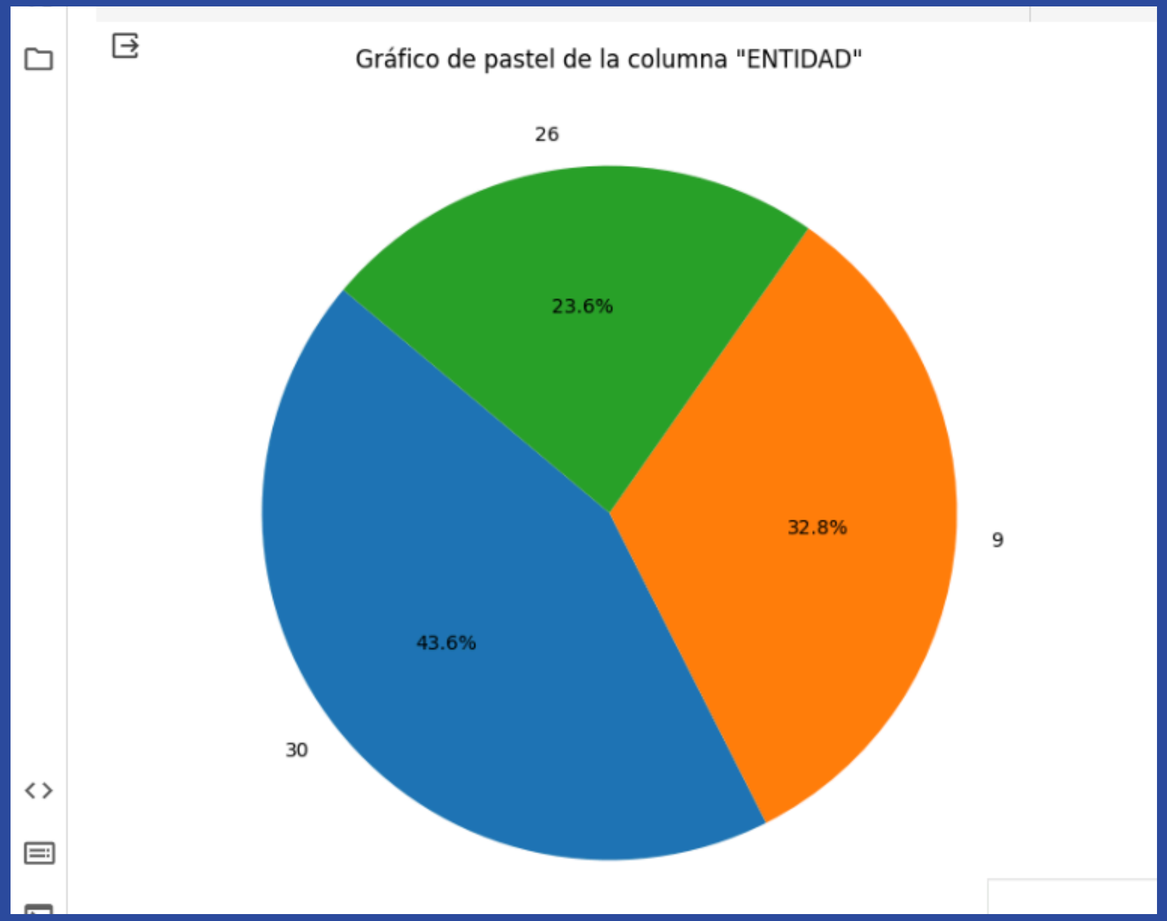
- ***Archivos y Registros:***
  - *Datos consolidados de tres estados: Ciudad de México, Sonora, Veracruz.*
  - Dimensiones del dataframe : *237,913 filas y 230 columnas.*

Dimensiones del dataframe limpio: 155,527 filas y 12 columnas.

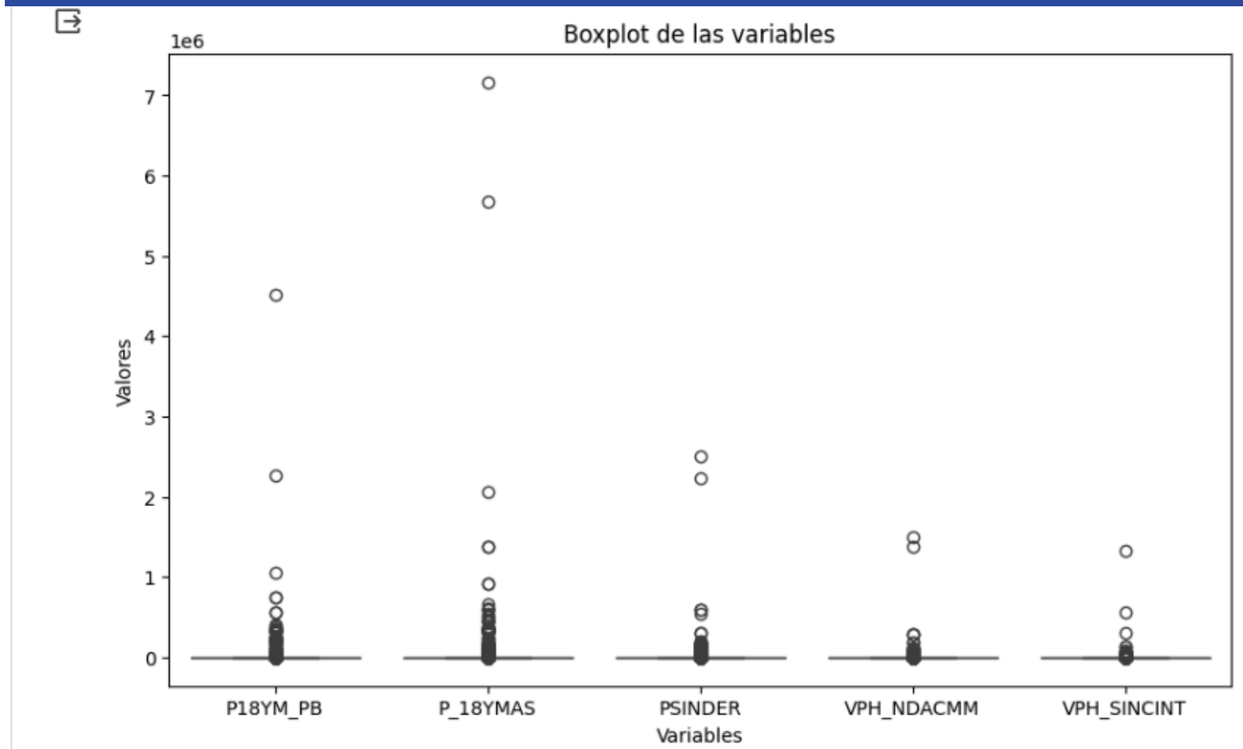
- ***Columnas Relevantes:***
  - **ENTIDAD**
  - **MUN**
  - **NOM_MUN**
  - **LOC**
  - **NOM_LOC**
  - **AGEB**
  - **MZA**
  - **P18YM_PB**
  - **P_18YMAS**
  - **PSINDER**
  - **VPH_NDACMM**
  - **VPH_SINCINT**

EXPLORACIÓN DE DATOS

DATOS CUALITATIVOS

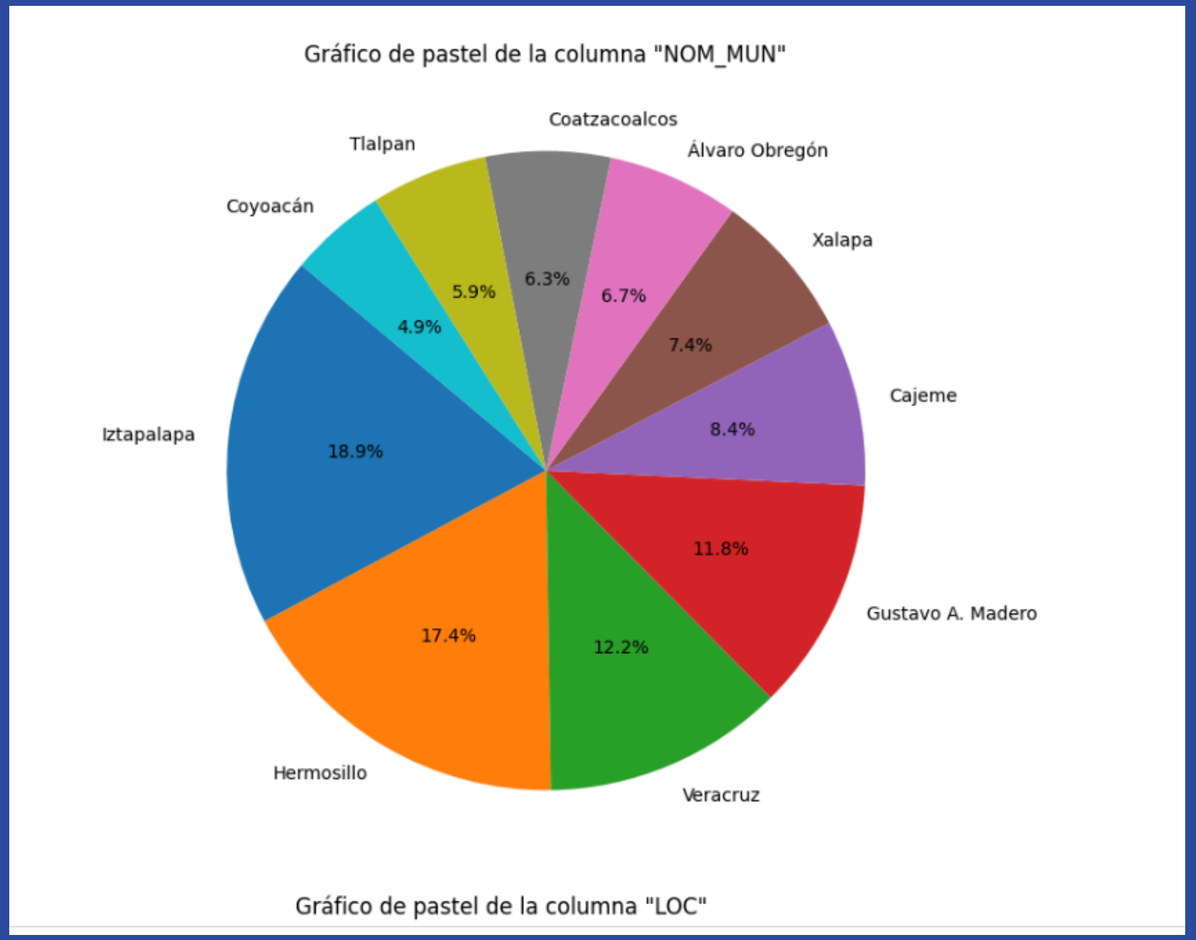


DATOS CUANTITATIVOS

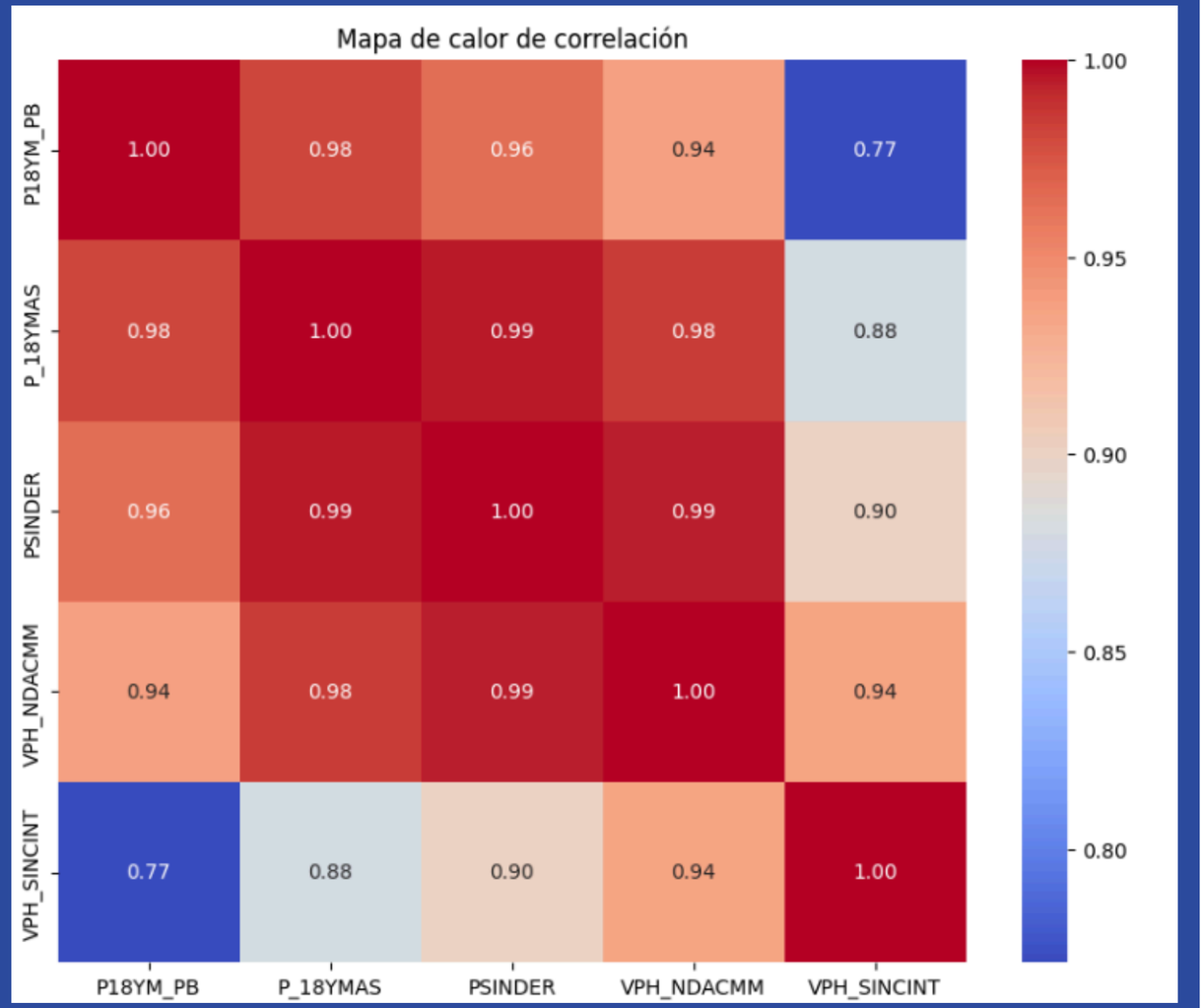


EXPLORACIÓN DE DATOS

DATOS CUALITATIVOS



DATOS CUANTITATIVOS



MODELO DE APRENDIZAJE SUPERVISADO

Se utilizaron tres modelos de regresión de la librería ScikitLearn:

- DecisionTreeRegressor
- KNeighborsRegressor
- RandomForestRegressor

MEDOTOLOGÍA

- Entrenamiento/Prueba
- GridSearchCV
- R2 Score

Tipo de Modelo	Pedro Soto Juárez	Francisco Obregón Gaxiola	Sergio Alejandro Esparza González	Marcos Renato Aquino	Arath Mendivil Mora	Carlos San Pedro
Random Forest	max_depth:20,min_samples_split:2,n_estimators:20 Coeficiente de determinación (R^2): 0.9012	max_depth: 20, min_samples_split: 2, n_estimator: 5 (R^2): 0.9881	max_depth: 20, min_samples_split: 2, n_estimator: 20 0.8107	max_depth:10,min_samples_split:2,n_estimators:20 (R^2): 0.98449	max_depth:10,min_samples_split:2,n_estimators: 20 (R^2): 0.7596	max_depth:10,min_samples_split:2,n_estimators:20 (R^2): 0.9958
K Neighbors	: k=2 (R^2): 0.4384	: k=2 (R^2): 0.4384	k=4 (R^2): 0.8200	k=2 (R^2: 0.4384	k=2 (R^2):0.7595	k=2 (R^2): 0.4384
Decision Trees	Max depth:10 y min_samples_split:2 0.4893 de coeficiente de determinación(R^2)	Max depth:20 y min_samples_split:2 0.4911	Max_depth: 10, min_samples_split: 2 0.7125	Max depth:5 y min_samples_split:2 0.4910	Max depth:20 y min_samples_split: 2 0.7596	Max depth:10 y min_samples_split:2 0.4910

SELECCIÓN Y DESPLIEGUE DEL MODELO

Modelo Seleccionado:

- Random Forest

Aplicación del modelo:

- Optimización de recursos: Utilizar el modelo para identificar áreas geográficas o grupos de población con una mayor propensión a no completar la educación post-básica, lo que permitiría dirigir los recursos y las intervenciones hacia donde sean más necesarios.

CONCLUSIONES

El estudio confirmó la viabilidad de usar modelos de regresión para predecir la educación post básica y las condiciones de vida asociadas. El Random Forest, con su capacidad para manejar múltiples variables y capturar complejidades no lineales, se estableció como el modelo más prometedor. Gracias a esto, se puede implementar en distintos municipios una forma de optimizar los recursos para que así más jóvenes mexicanos logren terminar sus estudios de preparatoria

ACTIVIDADES POR REALIZAR

- *Monitorear y actualizar el modelo regularmente con nuevos conjuntos de datos para mantener su relevancia.*
- *Identificar áreas geográficas o grupos de población con una mayor propensión a no completar la educación post-básica.*

RECOMENDACIONES

- *Promover este modelo en los distintos Estados de México para así optimizar el manejo de recursos donde es más necesario.*